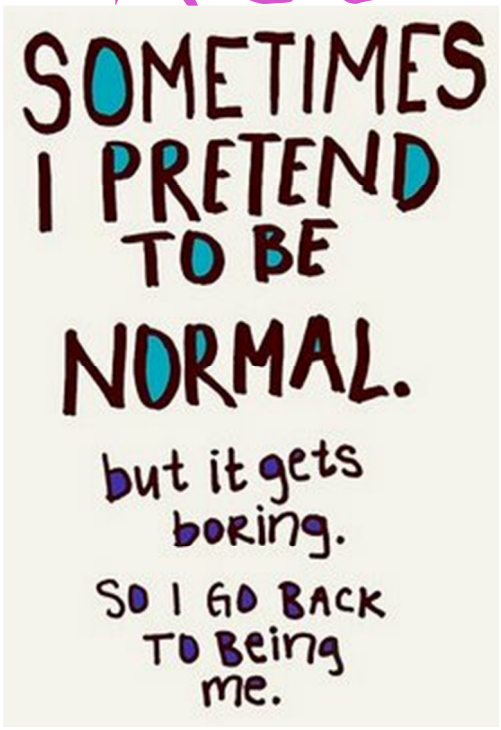
THE NORMAL DISTRIBUTION AND CORRELATION

HSC



General Maths

NAME: _____

HSC CAPACITY MATRIX – GENERAL MATHEMATICS

TOPIC: Data Analysis – Normal Distribution (DA6) & Correlation (DA7) **3 weeks**

CONTENT	CAPACITY BREAKDOWN!	DONE IT!!!!	GOT IT!!!!!	ON MY WAY!	WORKING ON IT!	HELP!!!!
1. Describing z-scores and using the formula to calculate z-scores.	Ex 11A					
2. Comparison of z-scores	Ex 11B					
3. Distribution of scores	Ex 11C					
4. Scatterplots	Ex 12A					
5. Median Regression lines	Ex 12B Q 2, 3, 4 & 12C					
6. Correlation	Ex 12D					

Your say!

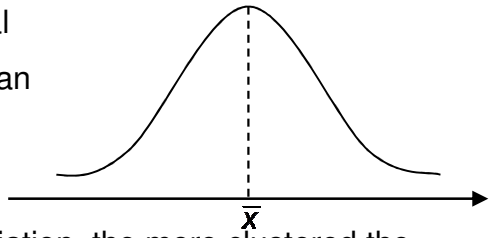
What was the most important thing you learned?_____

What was something new you learnt? _____

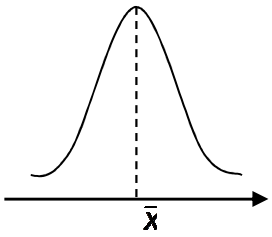
What part(s) of this topic will you need to work on? _____

z-scores

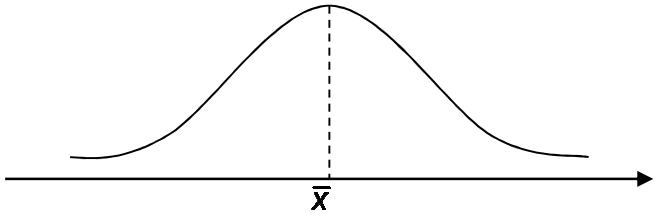
A **normal distribution** is a data set that is symmetrical about the mean. It is bell shaped and the mean, median and mode all take the same value.



The **spread** of the normal distribution will depend on the standard deviation. The lower the standard deviation, the more clustered the scores will be around the mean.



Normal distribution with a small standard deviation.



Normal distribution with a large standard deviation.

A **z-score (standardised score)** is used to measure the position of a score in a data set relative to the mean. It measures the distance from the mean in terms of the standard deviation.

- A **z-score of 0** indicates that the score obtained is **EQUAL** to the mean;
- A **negative z-score** indicates that the score obtained is **BELOW** the mean;
- A **positive z-score** indicates that the score obtained is **ABOVE** the mean;

A score that is exactly one standard deviation above the mean has a z-score of 1;
A score that is exactly one standard deviation below the mean has a z-score of -1

- To **calculate a z-score**:

$$z = \frac{x - \bar{x}}{s}$$

where x = score, $\bar{x}$ = mean and s = standard deviation.

- When examining z-scores, care must be taken to use the appropriate value for the standard deviation. If examining a population, the population standard deviation (σ_n) should be used. If a sample has been taken, the sample standard deviation (σ_{n-1}) or s_n should be used.

eg In an IQ test, the mean IQ is 100 and the standard deviation is 15. Student A's test results give an IQ of 130. Calculate this as a z-score and comment on Student A's result.

eg A sample of professional basketball players gives the mean height as 192 cm with a standard deviation of 12 cm.

a) Comment on the heights of the players;
b) Fred is 183 cm tall, Calculate Fred's height as a z-score.

eg To obtain the average number of hours study completed by Yr 12 students per week, Kim surveys 20 students and obtains the following results:

12	18	15	14	9	10	13	12	18	25
15	10	3	21	11	12	14	16	17	20

a) Calculate the mean and standard deviation, Correct to two decimal places;
b) Alex completes 16 hours of study each week. Express this as a z-score, correct to 2 decimal places.

Comparison of z-scores

- *Scores can be compared by their z-scores as they compare the score with the mean and standard deviation.*
- *Consider the question carefully – you will need to ascertain if the higher or lower z-score is a better outcome.*

eg Naomi scored 82 in her Classical Greek exam and 78 in her Latin for Beginners exam. In Greek, the mean was 62 and deviation of 10; while for Latin the mean was 66 and a deviation of 5.

a) Give both results as a standardised score;
b) Which is the better result – justify your answer.

eg In international swimming the mean time for the men's 100m free is 50.46s with a deviation of 0.6s. For the 200m free, the mean time is 1min 51.4s with a standard deviation of 1.4s.

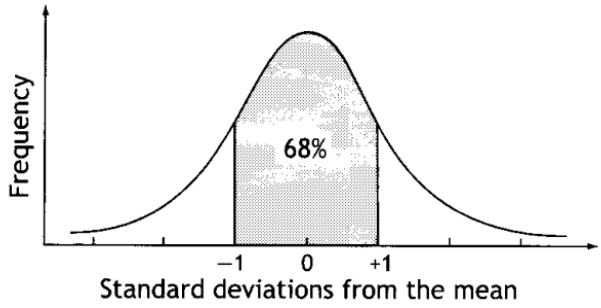
Bryce's best time is 49.92s for 100m and 1 min 49.3s for 200m. At a competition, Bryce can enter only one of these events. Which event should he enter and justify your answer.

DISTRIBUTION OF SCORES

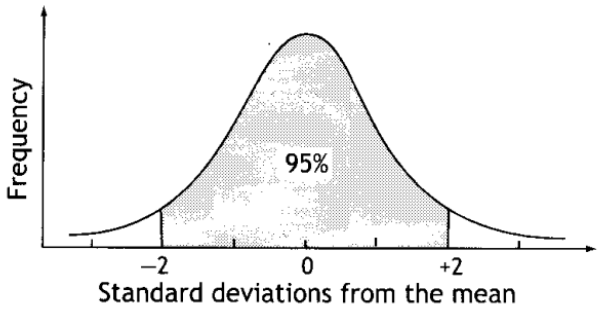
In any normal distribution, the percentage of scores that lie within a certain number of standard deviations of the mean is always the same, provided that the sample is large enough.

REMEMBER: the mean, mode and median of a normal distribution are the same.

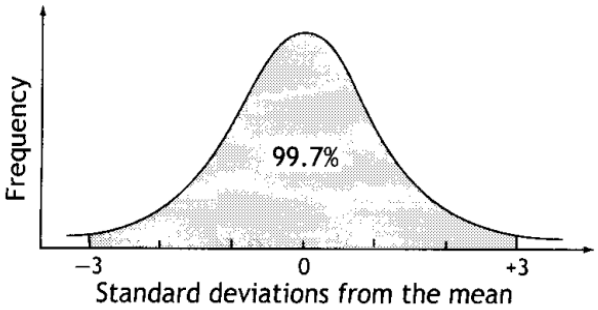
Approximately 68% of the population lies within one standard deviation of the mean ie $\bar{x} - 1\sigma n$ and $\bar{x} + 1\sigma n$
ie ± 1 z



Approximately 95% of the population lies within two standard deviations of the mean, ie
$\bar{x} - 2\sigma n$ and $\bar{x} + 2\sigma n$
ie ± 2 z



Approximately 99.7% of the population lies within three standard deviations of the mean, ie
$\bar{x} - 3\sigma n$ and $\bar{x} + 3\sigma n$
ie ± 3 z



From this we can deduce that:

- Any score will **almost certainly** lie within 3 z-scores (or standard deviations) of the mean;
- Any score will **very probably** lie within 2 z-scores (or standard deviations) of the mean.

eg the heights of 300 students are normally distributed. The mean height is 173 cm and the standard deviation is 3 cm. Find the number of students whose heights are between 170 cm and 176cm.

eg A machine produces rods of mean diameter 7.000 cm and standard deviation 0.030 cm. Within what interval will the diameters almost certainly lie?

eg Consider the rods described in the previous example. If 3 rods in a batch of 100 are found to have diameters of 7.098 cm, what can we conclude?

eg The scores obtained on a commonly used IQ test can be assumed to be normally distributed with a mean of 100 and a standard deviation of 15. Approximately what percentage of the distribution lies:

a) between 85 and 115?
b) Between 70 and 130?
c) Between 55 and 145?

eg In an exam the mean was 60 and standard deviation of 12. What percentage of candidates in the exam scored above 84?

Scatterplots

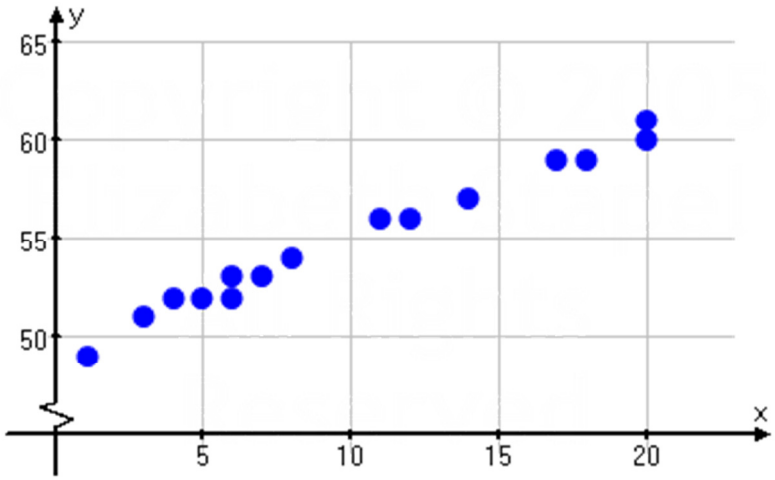
http://www.purplemath.com/modules/scattreg.htm

Real life is messy, so it is expected that measurements taken from real life will be messy as well. And when you graph these measurements of real life, it is expected that the dots won't line up exactly in a nice neat line, but will instead form a scattering of dots which, at best, might suggest a nice neat line. These dots are called a scatterplot.

- **Create a scatterplot from the following data:**

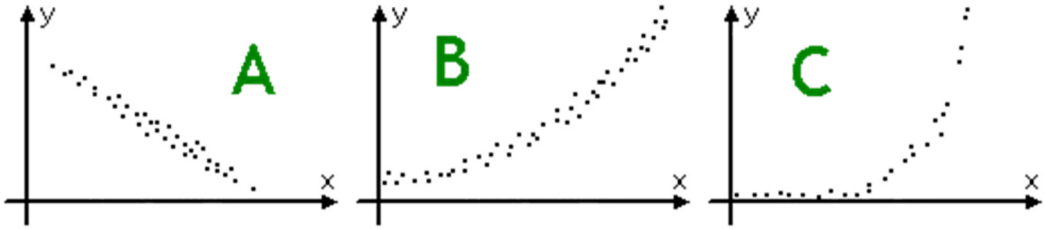
  (1, 49), (3, 51), (4, 52), (6, 52), (6, 53), (7, 53), (8, 54), (11, 56), (12, 56), (14, 57), (14, 58), (17, 59), (18, 59), (20, 60), (20, 61)

  One of the first things I have to do when graphing these points is figure out what my axis scale values are going to be. If I try doing an axis system with the "standard" –10 to 10 values, none of the above points will even show up on my graph. As is common with these sorts of data sets, all the x- and y-values are positive, so I only really need scales for the first quadrant. The y-values are much larger than the x-values, but instead of squeezing all the y-values together, I'll spread them out (so I can see them better) by using an interrupted scale.



Usually you'll be working with scatterplots where the dots line up in some sort of vaguely straight row. But you shouldn't expect everything to line up nice and neat, especially in "real life" (like, for instance, in a physics lab). And sometimes you'll need to pick a different sort of equation as a model, because the dots line up, but not in a straight line.

- **Tell which sort of equation you think would best model the data in the following scatterplots, and why.**



- **Graph A:** The dots look like they line up fairly straight, so a **linear model** would probably work well.
- **Graph B:** The dots here do line up, but as more of a curvy line. A **quadratic model** might work better.
- **Graph C:** The dots are very close to the *x*-axis, and then they shoot up, so an **exponential or power-function model** might work better here.
- **Note that both Graphs B and C can be referred to as nonlinear models.**

In general, expect only to need to recognise linear (straight-line) versus quadratic (curvy-line) models, and never anything that you haven't already covered in class.

NOTE:

A relationship between the variables exists if one increases as the other increases OR if one decreases as the other increases.

Bivariate Data and the Line of Best fit (Median Regression Line)

http://illuminations.nctm.org/ActivityDetail.aspx?ID=146

http://illuminations.nctm.org/ActivityDetail.aspx?ID=82

If there are two types of data in a study, they are called **BIVARIATE DATA.**

The process of fitting straight lines to bivariate data on a scatterplot enables us to analyse relationships between the data and possibly make predictions based on the given data set.

The line of best fit is called the **Median Regression line.**

ACTIVITY:

http://www.youtube.com/watch?v=YgBpCfQVfzE

Is there a relationship between the length of your middle finger and the width of your hand?

1. Measure the length of your middle finger and the width of your hand and write them down.
2. Enter the data onto the interactive site graph and determine if it is linear.
3. Determine the line of best fit.
4. Calculate the equation of the line:
   a. **Create a right triangle using the line as the hypotenuse (try to find two points on the line as the endpoints of the hypotenuse);**
   b. **Calculate the vertical and horizontal lengths;**
   c. **Then calculate the gradient** $m = \frac{rise}{run}$ **Remember that if the line leans to the left, the gradient is negative, leans to the right, the gradient is positive**
   d. **Find the point where the line cuts the y-axis and determine the y-value (b);**
   e. **Put into the form** $m = mx + b$
5. From your equation, calculate the length of a middle finger on the hand that measures 13.8cm across.
6. From your equation, calculate the width of a hand that has a middle finger length of 11.1cm.
7. From the line of best fit, comment on the relationship between the two variables.

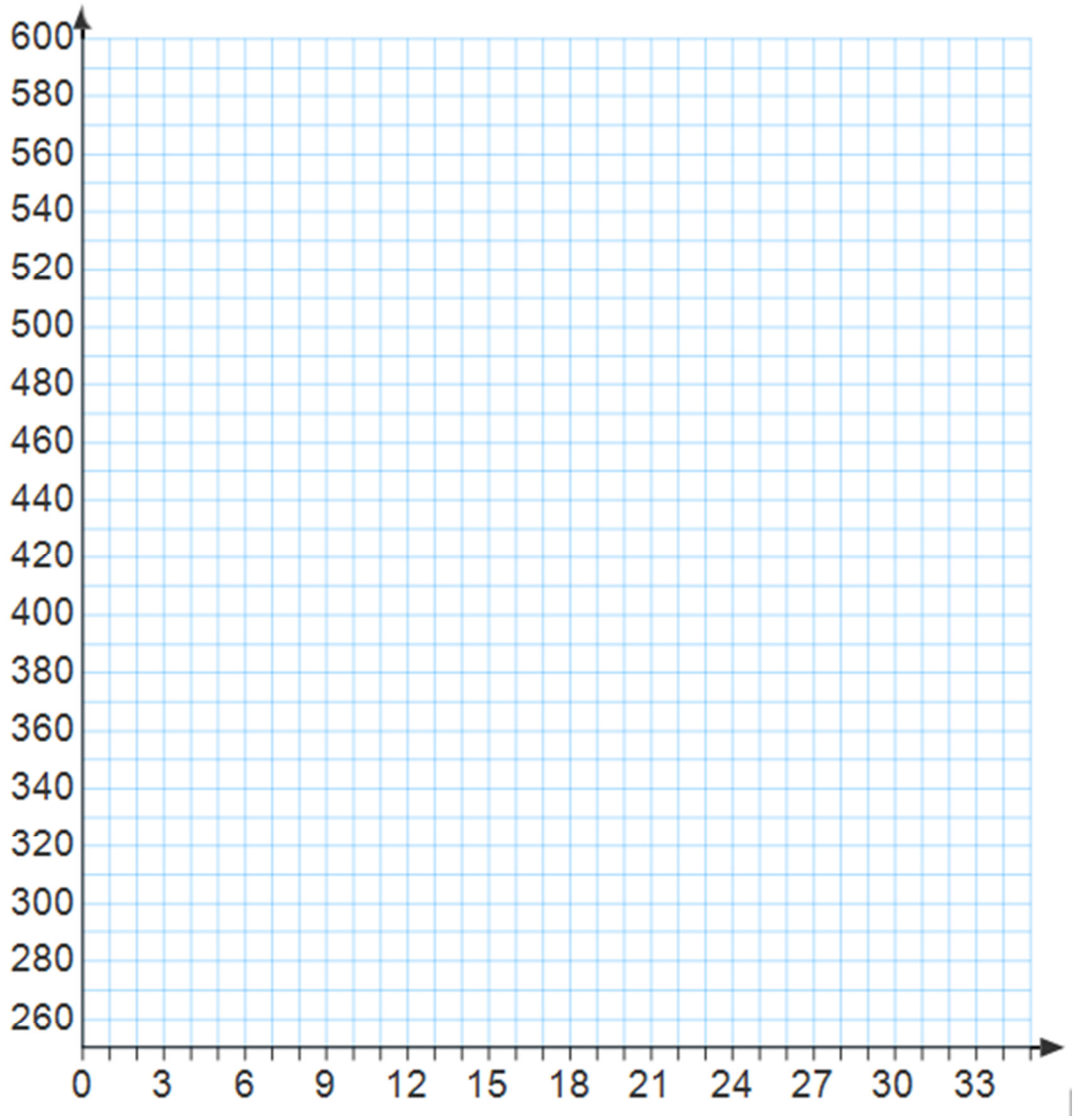
ACTIVITY:

Is there a relationship between the fat Grams and the total calories in fast food?

Sandwich	Total Fat (g)	Total Calories
Hamburger	9	260
Cheeseburger	13	320
Quarter Pounder	21	420
Quarter Pounder with Cheese	30	530
Big Mac	31	560
Arch Sandwich Special	31	550
Arch Special with Bacon	34	590
Crispy Chicken	25	500
Fish Fillet	28	560
Grilled Chicken	20	440
Grilled Chicken Light	5	300

1. Name the two two types of data in this sample.

2. On the graph paper below, create a scatterplot to represent the data above.



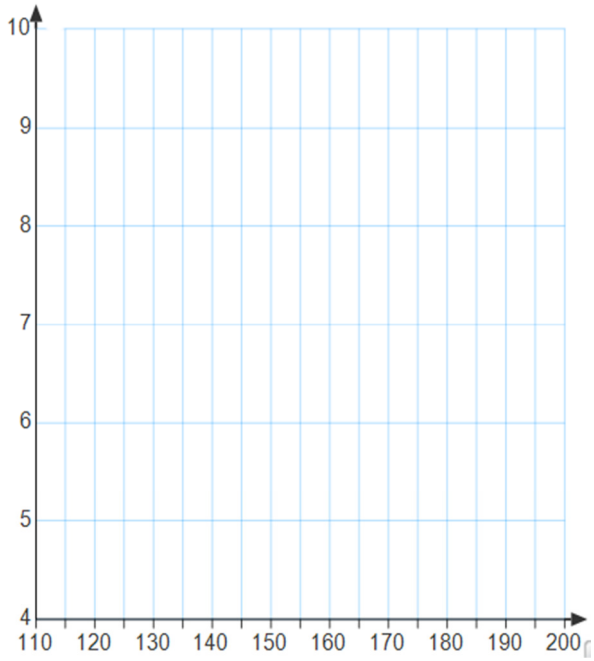
3. Determine if the data is a linear representation – if so, draw in a line of best fit.
4. Hence find the equation of the line.

5. Comment on the relationship of the data.

1 The following table lists the population (to the nearest ten thousand) and the number of primary schools in each of 11 municipalities.

Population (×1000)	No. of primary schools
110	4
130	4
130	6
140	5
150	6
160	8
170	6
170	7
180	8
180	9
190	8

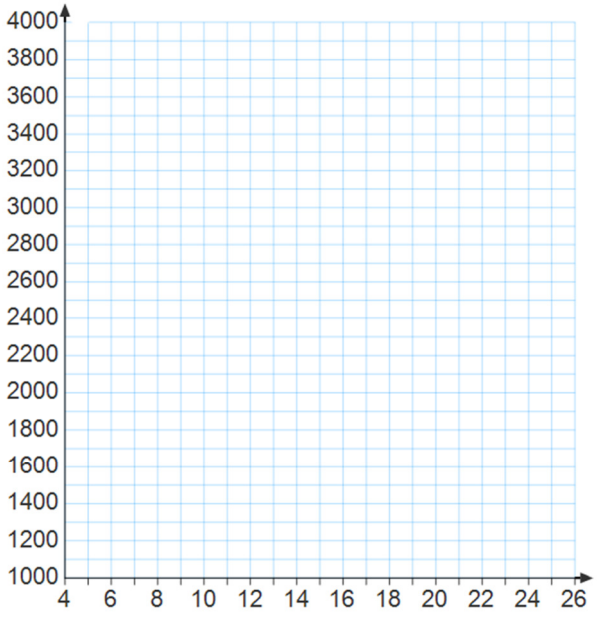
Construct a scatterplot for the data.



2 The table below contains data for the time taken to do a paving job and the cost of the job.

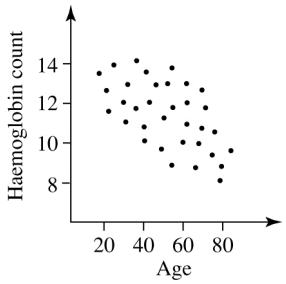
Time taken (hours)	Cost of job ($)
5	1000
7	1000
5	1500
8	1200
10	2000
13	2500
15	2800
20	3200
18	2800
25	4000
23	3000

Construct a scatterplot of the data.

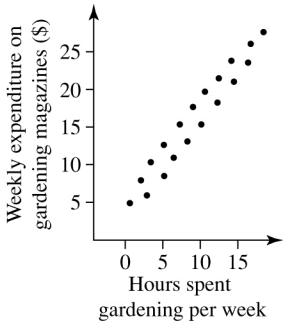


3 For each of the scatterplots below state if a relationship exists and if so, state if the relationship is linear.

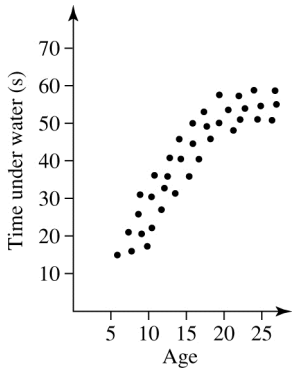
(a)



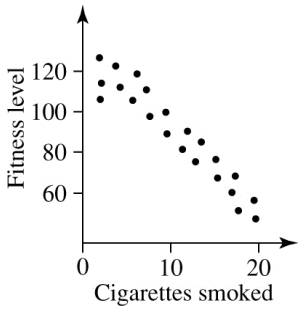
(b)



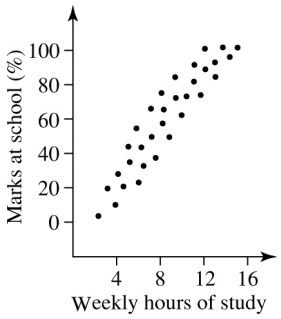
(c)



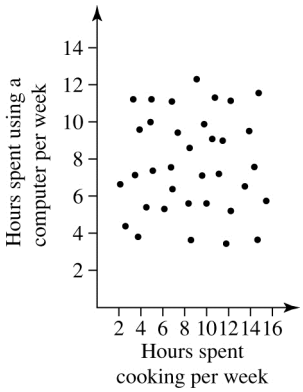
(d)



(e)

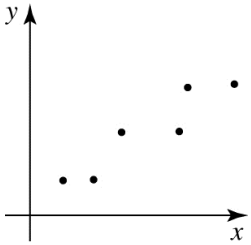


(f)

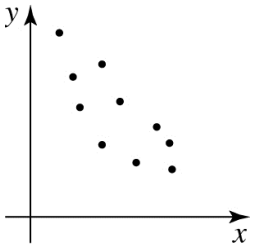


4 Fit a straight line to the data in the following scatterplots using the equal-number-of-points method.

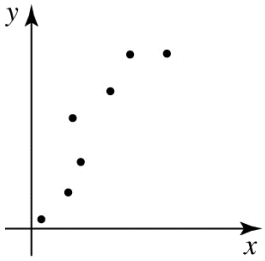
(a)



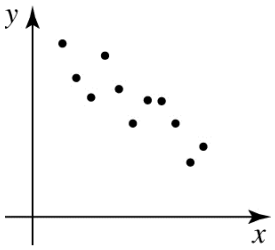
(b)



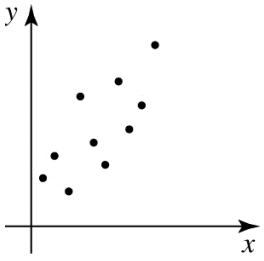
(c)



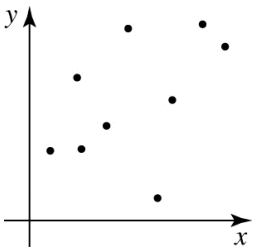
(d)



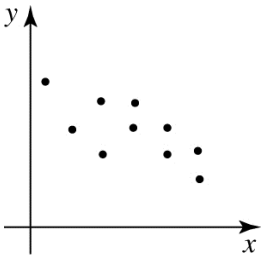
(e)



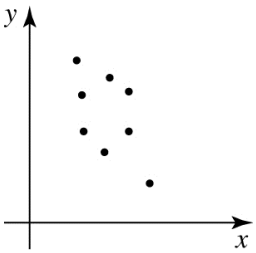
(f)



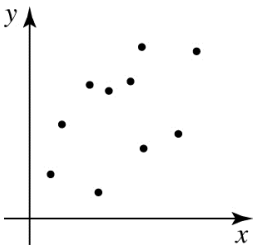
(g)



(h)



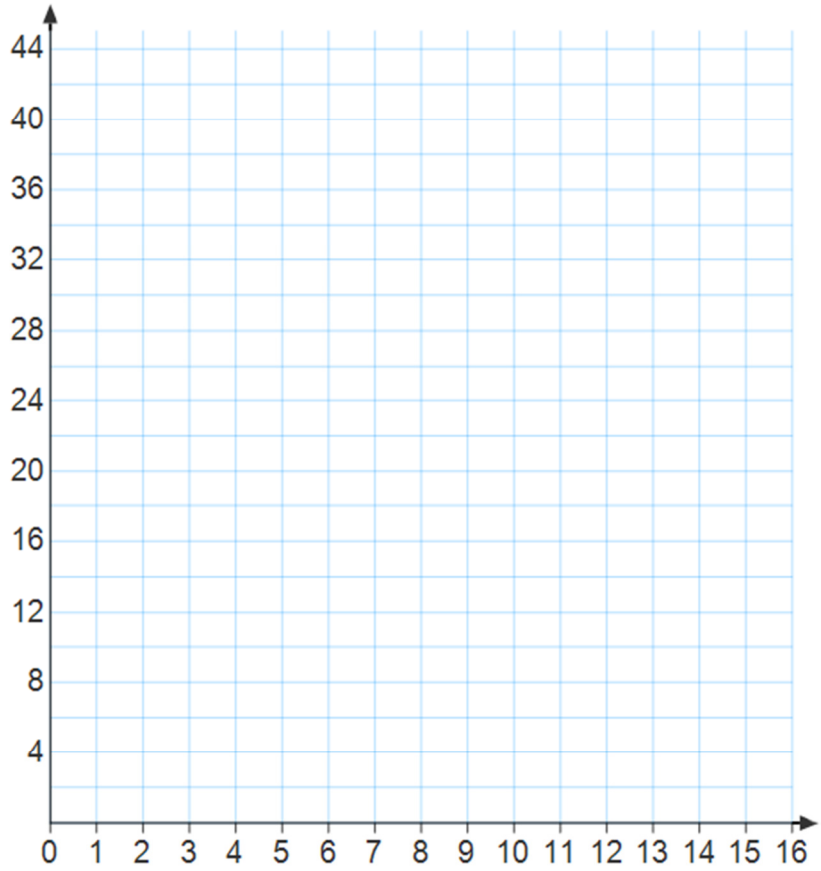
(f)



5 The table below contains data giving the time taken to engineer a finished product from the raw recording (e.g. a song) and the length of the finished product.

Hours spent in engineering studio	Finished length of recording (min)
1	3
2	4
3	10
4	12
5	20
6	16
7	18
8	25
9	30
10	28
11	35
12	36
13	39
14	42
15	45

(a) Construct a scatterplot for these data.



(b) Comment on whether a relationship exists between the time spent engineering and the length of the finished recording.

(c) If a relationship exists, state if the relationship is linear.

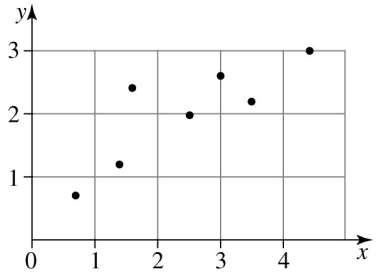
6 Find the approximate regression line equation for the data in the table below using the 3-median method.

x	2	3	3	4	4	5	5	7
y	1	3	5	3	6	6	7	8

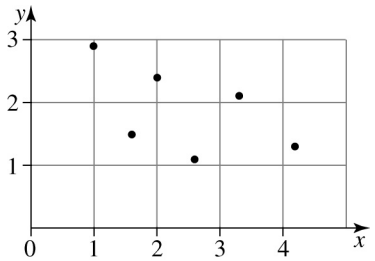
---

7 By eye, fit a straight line to the data in the following graphs using the equal-number-of-points method.

(a)



(b)



---

8 Find the approximate regression line equation for the data in the table below using the 3-median method.

x	5	10	20	20	30	40	50	55
y	80	60	50	70	40	55	40	30

---

9 Suppose that a line of best fit has a gradient of 3.5 and passes through the point (−6, 13). Find:

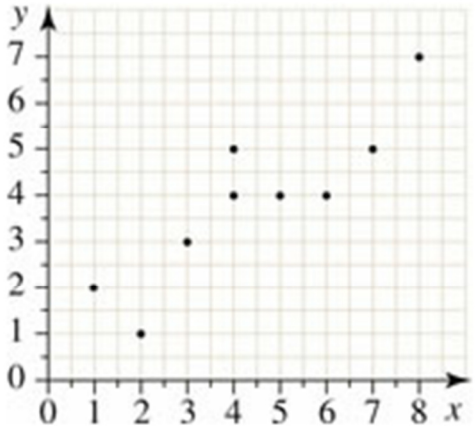
(a) the equation of the line

(b) the value of y when $x = 16$.

Fitting a straight line – the 3-median regression line

For greater accuracy you need closer analysis of the data.
One method is called the 3-median regression line which follows these steps:

STEP 1 Plot points on a scattergram.



STEP 2 Divide the points into 3 groups using vertical divisions following these rules:

(a) Divide into 3 equal groups, or
(b) If there is one extra point, add to the middle group;
(c) If there are 2 extra points, put 1 point in each of the outer groups

STEP 3 Find the median point of each group and mark in order:

(a) The left group is the lower group and it median is (x_L, y_L)
(b) The middle group's median is labelled (x_M, y_M)
(c) The right group is the upper groups and its median is labelled $(x_{U,} y_U)$

STEP 4 Draw in the line of best fit. Place a straight edge so that it passes through the lower and upper medians. Move the edge a third of the way toward the middle group median WHILE MAINTAINING THE SLOPE. Draw the median regression line.

STEP 5 Find the equation of the line.